Optimization of the Asymptotic Property of Mutual Learning Involving an Integration Mechanism of Ensemble Learning

Kazuyuki Hara[1] *, Takahiro Yamada[2]

[1]Tokyo Metropolitan College of Industrial Technology
Higashi-oi 1-10-40, Shinagawa-ku, Tokyo 140-0011.
[2]Toyohashi University of Technology
1-1, Hibarigaoka, Tempaku, Toyohashi, Aichi, 441-8580.

Abstruct– We propose an optimization method of mutual learning which converges into the identical state of optimum ensemble learning within the framework of on-line learning, and have analyzed its asymptotic property through the statistical mechanics method.The proposed model consists of two learning steps: two students independently learn from a teacher, and then the students learn from each other through the mutual learning. In mutual learning, students learn from each other and the generalization error is improved even if the teacher has not taken part in the mutual learning. However, in the case of different initial overlaps(direction cosine) between teacher and students, a student with a larger initial overlap tends to have a larger generalization error than that of before the mutual learning. To overcome this problem, our proposed optimization method of mutual learning optimizes the step sizes of two students to minimize the asymptotic property of the generalization error. Consequently, the optimized mutual learning converges to a generalization error identical to that of the optimal ensemble learning. In addition, we show the relationship between the optimum step size of the mutual learning and the integration mechanism of the ensemble learning.

Keywords– mutual learning, learning step size, on-line learning, linear perceptron, statistical mechanics

1 Introduction

As a model involving the interaction between students, Kinzel proposed mutual learning within the framework of on-line learning[9, 10, 11]. Kinzel's model employs two students, and a student learns with the other student acting as a teacher. The target of his model is to obtain the same networks through the learning. On the other hand, ensemble learning algorithms, such as bagging[1]

*E-mail:hara@tokyo-tmct.ac.jp

and Ada-boost[2], try to improve upon the performance of a weak learning machine by using many weak learning machines; such learning algorithms have recently received considerable attention. We have noted, however, that the mechanism of integrating the outputs of many weak learners in ensemble learning is similar to that of obtaining the same networks through mutual learning.

From the point of view of the learning problem, how the student approaches the teacher is important. However, Kinzel[9, 10, 11] does not deal with the teacher-student relation since a teacher is not employed in his model. In contrast to Kinzel's model, we have proposed mutual learning between two students who learn from a teacher in advance[12]. In our previous work[12], we showed that the generalization error of the students becomes smaller through the mutual learning even if the teacher does not take part in the mutual learning. We also showed that a student with a larger initial overlap(direction cosine) for mutual learning transiently passes through a state of the optimum ensemble learning when the limit of the learning step size is zero.

In this paper, we propose a new mutual learning algorithm that uses a different learning step size for each student. We analyze the asymptotic property of the proposed learning algorithm through the statistical mechanics method, and propose an optimization method for the learning step size. By using the optimum learning step size, we can obtain the optimum asymptotic property of the generalization error through mutual learning. The proposed method is an expansion of our previous work[12].

In this paper, we assume that each teacher and student is a linear perceptron. An on-line learning[3] scheme is employed. In the proposed method, two students individually learn from a teacher during initial learning, and then they learn from each other during mutual learning. Therefore, we assume the overlaps between teacher and students are not zero at the initial state of mutual learning. In the mutual learning, each student learns from the other as the teacher. Since a teacher is not used in the mutual learning, we refer to a latent teacher in this paper.

In Section 2, we formulate latent teacher, student, and mutual learning algorithms. In Section 3, we derive differential equations of the order parameters that depict the dynamics of mutual learning. We employ different learning step sizes for each student. We then derive the generalization error by using the order parameters. In Section 4, we solve the differential equations with different learning step sizes, and then analyze the effect of the learning step size on the asymptotic property of the mutual learning. After that, we obtain the optimum ratio of the students' learning step sizes which realizes the minimum generalization error. Moreover, we discuss the relation between the learning step size of mutual learning and the integration mechanism of ensemble learning.

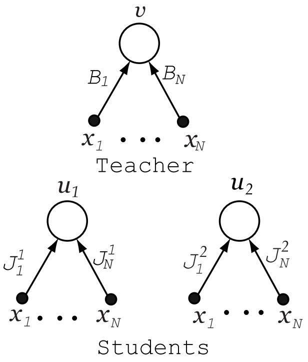


Figure 1: Network structure of latent teacher and student networks, all having the same network structure.

2 Formulation of mutual learning with a latent teacher

In this section, we formulate the latent teacher and student networks, and the mutual learning algorithms. We assume the latent teacher and student networks receive N-dimensional input $\boldsymbol{x}(m) = (x_1(m), \ldots, x_N(m))$ at the m-th learning iteration as shown in Fig. 1. Learning iteration m is ignored in the figure. The latent teacher network is a linear perceptron, and the student networks are two linear perceptrons. We also assume that the elements $x_i(m)$ of the independently drawn input $\boldsymbol{x}(m)$ are uncorrelated random variables with zero mean and $1/N$ variance; that is, the elements are drawn from a probability distribution $P(\boldsymbol{x})$. In this paper, the thermodynamic limit of $N \to \infty$ is assumed. The size of input vector $|\boldsymbol{x}|$ then becomes one.

$$\langle x_i \rangle = 0, \quad \langle (x_i)^2 \rangle = \frac{1}{N}, \quad |\boldsymbol{x}| = 1, \tag{1}$$

where $\langle \cdots \rangle$ denotes average, and $| \cdot |$ denotes the norm of a vector.

The latent teacher network is a linear perceptron, and is not subject to training. Thus, the weight vector is fixed in the learning process. The output of the latent teacher $v(m)$ for N-dimensional input $\boldsymbol{x}(m) = (x_1(m), x_2(m), \ldots, x_N(m))$ at the m-th learning iteration is

$$v(m) = \sum_{i=1}^{N} B_i x_i(m) = \boldsymbol{B} \cdot \boldsymbol{x}(m), \tag{2}$$

$$\boldsymbol{B} = (B_1, B_2, \ldots, B_N), \tag{3}$$

where latent teacher weight vector $\boldsymbol{B}$ is an N-dimensional vector like the input vector, and each element B_i of the latent teacher weight vector $\boldsymbol{B}$ is drawn from a probability distribution of zero mean and unit variance. Assuming the thermodynamic limit of $N \to \infty$, the size of latent teacher weight vector $|\boldsymbol{B}|$ becomes $\sqrt{N}$.

$$\langle B_i \rangle = 0, \quad \langle (B_i)^2 \rangle = 1, \quad |\boldsymbol{B}| = \sqrt{N}. \tag{4}$$

The output distribution for the latent teacher $P(v)$ follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit of $N \to \infty$.

The two linear perceptrons are used as student networks that compose the mutual learning machine. Each student network has the same architecture as the latent teacher network. Each element of $\boldsymbol{J}^k(0)$ which is the initial value of the k-th student weight vector $\boldsymbol{J}^k$ is drawn from a probability distribution of zero mean and unit variance. The norm of the initial student vector $|\boldsymbol{J}^k(0)|$ is $\sqrt{N}$ in the thermodynamic limit of $N \to \infty$,

$$\langle J_i^k(0) \rangle = 0, \ \ \langle (J_i^k(0))^2 \rangle = 1, \ \ |\boldsymbol{J}^k(0)| = \sqrt{N}. \tag{5}$$

The k-th student output $u_k(m)$ for the N-dimensional input $\boldsymbol{x}(m)$ is

$$u_k(m) = \sum_{i=1}^{N} J_i^k(m) x_i(m) = \boldsymbol{J}^k(m) \cdot \boldsymbol{x}(m), \tag{6}$$

$$\boldsymbol{J}^k(m) = (J_1^k, J_2^k, \ldots, J_N^k). \tag{7}$$

Generally, the norm of student weight vector $|\boldsymbol{J}^k(m)|$ changes as the time step proceeds. Therefore, the ratio l_k of the norm to $\sqrt{N}$ is considered and is called the length of student weight vector $\boldsymbol{J}^k$. The norm at the m-th iteration is $l_k(m)\sqrt{N}$, and the size of $l_k(m)$ is $O(1)$.

$$|\boldsymbol{J}^k(m)| = l_k(m)\sqrt{N} \tag{8}$$

The distribution of the output of the k-th student $P(u_k)$ follows a Gaussian distribution of zero mean and l_k^2 variance in the thermodynamic limit of $N \to \infty$.

Next, we formulate the learning algorithm. After the students learn from a latent teacher, mutual learning is carried out. The learning equation of the mutual learning is

$$\boldsymbol{J}^k(m+1) = \boldsymbol{J}^k(m) + \eta_k \Big(\ u_{k'}(m) - u_k(m) \ \Big) \boldsymbol{x}(m), \tag{9}$$

where k is 1 or 2 and $k \neq k'$. m denotes the iteration number. Equation (9) shows that mutual learning is carried out between two students. Therefore, the teacher used in the initial learning is called a latent latent teacher. We use the gradient descent algorithm in this paper, while another algorithm was used in Kinzel's work [9]. When the interaction between students is introduced, the performance of students may be improved if they exchange knowledge that each student has acquired from the latent teacher in the initial learning. In other words, two students approach each other through mutual learning, and tend to move towards the middle of the initial weight vectors. This tendency is similar to the integration mechanism of ensemble learning, so mutual learning may mimic this mechanism.

3 Theory

In this section, we first derive the differential equations of two order parameters which depict the behavior of mutual learning. After that, we derive an auxiliary order parameter which depicts the relationship between the latent teacher and students. We then rewrite the generalization error using these order parameters.

We first derive the differential equation of the length of the student weight vector l_k. l_k is the first order parameter of the system. We modify the length of the student weight vector in Eq. (8) as $\boldsymbol{J}^k \cdot \boldsymbol{J}^k = Nl_k^2$. To obtain a time dependent differential equation of l_k, we square both sides of Eq. (9). We then average the term of the equation using the distribution of $P(u_k, u_{k'})$. Note that $\boldsymbol{x}$ and $\boldsymbol{J}^k$ are random variables, so the equation becomes a random recurrence formula. We formulate the size of the weight vectors to be $O(N)$, and the size of input $\boldsymbol{x}$ is $O(1)$, so the length of the student weight vector has a self-averaging property. Here, we rewrite m as $m = Nt$, and represent the learning process using continuous time t in the thermodynamic limit of $N \to \infty$. We then obtain the deterministic differential equation of l_k,

$$\frac{dl_k^2}{dt} = (\eta_k^2 - 2\eta_k)l_k^2 + \eta_k^2 l_{k'}^2 - 2(\eta_k^2 - \eta_k)Q. \tag{10}$$

Here, k is 1 or 2, and $k \neq k'$. In this equation, $Q = ql_k l_{k'}$ and q is the overlap between $\boldsymbol{J}^k$ and $\boldsymbol{J}^{k'}$, defined as

$$q = \frac{\boldsymbol{J}_k \cdot \boldsymbol{J}_{k'}}{|\boldsymbol{J}^k|\ |\boldsymbol{J}^{k'}|} = \frac{\boldsymbol{J}^k \cdot \boldsymbol{J}^{k'}}{Nl_k l_{k'}}, \tag{11}$$

and q is the second order parameter of the system. The overlap q also has a self-averaging property, so we can derive the differential equation in the thermodynamic limit of $N \to \infty$. The differential equation is derived by calculating the product of the learning equation (Eq. (9)) for $\boldsymbol{J}^k$ and $\boldsymbol{J}^{k'}$, and we then average the term of the equation using the distribution of $P(u_k, u_{k'})$. After that, we obtain the deterministic differential equation as

$$\frac{dQ}{dt} = (\eta_2 - \eta_1\eta_2)l_1^2 + (\eta_1 - \eta_1\eta_2)l_2^2 - (\eta_1 + \eta_2 - 2\eta_1\eta_2)Q. \tag{12}$$

Equations (10) and (12) form closed differential equations.

The analytical solutions of the length of the student l_k and the overlap between students Q are given by

$$l_k^2(t) = -A_1\frac{\eta_k}{\eta_{k'}}\exp(-(\eta_1 + \eta_2)(2 - (\eta_1 + \eta_2))t) + (-1)^k 2A_2\frac{\eta_k}{\eta_2 - \eta_1}\exp(-(\eta_1 + \eta_2)t) + A_3, \tag{13}$$

$$Q(t) = A_1 \exp(-(\eta_1 + \eta_2)(2 - (\eta_1 + \eta_2))t) + A_2 \exp(-(\eta_1 + \eta_2)t) + A_3, \tag{14}$$

where

$$A_1 = -\frac{\eta_1\eta_2(l_1^2(0) + l_2^2(0) - 2Q(0))}{(\eta_1 + \eta_2)^2}, \tag{15}$$

$$A_2 = -\frac{(\eta_2 - \eta_1)(\eta_2 l_1^2(0) - \eta_1 l_2^2(0) - (\eta_2 - \eta_1)Q(0))}{(\eta_1 + \eta_2)^2}, \tag{16}$$

$$A_3 = \frac{\eta_2^2 l_1^2(0) + \eta_1^2 l_2^2(0) + 2\eta_1\eta_2 Q(0)}{(\eta_1 + \eta_2)^2}. \tag{17}$$

$l_1(0)$ is the initial condition of student 1, and $l_2(0)$ is that of student 2. $Q(0) = q(0)l(0)$, and $q(0)$ is the initial condition of the overlap between student 1 and student 2. From Eqs. (13) and (14), $l_k^2(t)$ and $Q(t)$ converge to finite values at $t \to \infty$ if $2 - (\eta_1 + \eta_2) > 0$ is satisfied. Then the convergence condition of $l_k^2(t)$ and $Q(t)$ is

$$\eta_1 + \eta_2 \geq 2. \tag{18}$$

To depict the behavior of mutual learning with a latent latent teacher, we have to obtain the differential equation of overlap R_k, which is a direction cosine between latent teacher weight vector $\boldsymbol{B}$ and the k-th student weight vector $\boldsymbol{J}^k$ defined by Eq. (19). We introduce R_k as the third order parameter of the system.

$$R_k = \frac{\boldsymbol{B}\cdot\boldsymbol{J}^k}{|\boldsymbol{B}|\ |\boldsymbol{J}^k|} = \frac{\boldsymbol{B}\cdot\boldsymbol{J}^k}{Nl_k} \tag{19}$$

For the sake of convenience, we write the overlap between the latent teacher weight vector and the student weight vector as r_k and $r_k = R_k l_k$. The differential equation of overlap r_k is derived by calculating the product of $\boldsymbol{B}$ and Eq. (9), and we then average the term of the equation using the distribution of $P(v, u_k, u_{k'})$. The overlap r_k also has a self-averaging property, and in the thermodynamic limit the deterministic differential equation of r_k is then obtained through a calculation similar to that used for l_k.

$$\frac{dr_k}{dt} = \eta_k(r_{k'} - r_k) \tag{20}$$

The solution for overlap r_k is obtained by solving simultaneous differential equations of Eq. (20) for $k = 1$ and $k' = 2$, and for $k = 2$ and $k' = 1$.

$$r_k(t) = \frac{\eta_k(r_k(0) - r_{k'}(0))}{\eta_1 + \eta_2}\exp(-(\eta_1 + \eta_2)t) + \frac{\eta_2 r_1(0) + \eta_1 r_2(0)}{\eta_1 + \eta_2}, \tag{21}$$

where $r_k(0) = R_k(0)l(0)$, and $R_k(0)$ is the initial overlap between the latent teacher and the k-th student.

The squared error for the k-th student ϵ^k is then defined using the output of the latent teacher and that of the student as given in Eqs. (2) and (6), respectively.

$$\epsilon^k = \frac{1}{2}\Big(\boldsymbol{B}\cdot\boldsymbol{x} - \boldsymbol{J}^k\cdot\boldsymbol{x}\Big)^2 \tag{22}$$

The generalization error for the k-th student ϵ_g^k is given by the squared error ϵ^k in Eq. (22) averaged over the possible input $\boldsymbol{x}$ drawn from a Gaussian distribution $P(\boldsymbol{x})$ of zero mean and $1/N$ variance.

$$\begin{aligned} \epsilon_g^k &= \int d\boldsymbol{x} P(\boldsymbol{x})\ \epsilon^k \qquad (23) \\ &= \frac{1}{2}\int d\boldsymbol{x} P(\boldsymbol{x}) \Big(\boldsymbol{B}\cdot\boldsymbol{x} - \boldsymbol{J}^k\cdot\boldsymbol{x} \Big)^2 . \qquad (24) \end{aligned}$$

This calculation is the N-th Gaussian integral with $\boldsymbol{x}$ and it is hard to calculate. To overcome this difficulty, we employ coordinate transformation from $\boldsymbol{x}$ to v and u_k in Eqs. (2) and (6). Note that the distribution of the output of the students $P(u_k)$ follows a Gaussian distribution of zero mean and l_k^2 variance in the thermodynamic limit of $N \to \infty$. For the same reason, the output distribution for the latent teacher $P(v)$ follows a Gaussian distribution of zero mean and unit variance in the thermodynamic limit. Thus, the distribution $P(v, u_k)$ of latent teacher output v and the k-th student output u_k is

$$\begin{aligned} P(v, u_k) &= \frac{1}{2\pi\sqrt{|\Sigma|}} \exp\left[-\frac{(v, u_k)^T \Sigma^{-1}\ (v, u_k)}{2} \right] \qquad (25) \\ \Sigma &= \begin{pmatrix} 1 & r_k \\ r_k & l_k^2 \end{pmatrix} \qquad (26) \end{aligned}$$

Here, T denotes the transpose of a vector, r_k denotes $r_k = R_k l_k$, and R_k is the overlap between the latent teacher weight vector $\boldsymbol{B}$ and the student weight vector $\boldsymbol{J}^k$ defined by Eq. (19). Hence, by using this coordinate transformation, the generalization error in Eq. (24) can be rewritten as

$$\begin{aligned} \epsilon_g^k &= \frac{1}{2}\int dv du_k (v - u_k)^2 \qquad (27) \\ &= \frac{1}{2}(1 - 2r_k + l_k^2). \qquad (28) \end{aligned}$$

Consequently, we calculate the dynamics of the generalization error by substituting the time step value of $l_k(t)$, $Q(t)$, and $r_k(t)$ into Eq. (28).

$$\begin{aligned} \epsilon_g^k = \frac{1}{2}\Bigg\{ & 1 - 2\frac{\eta_k(r_k(0) - r_{k'}(0))}{\eta_1+\eta_2}\exp(-(\eta_1+\eta_2)t) - 2\frac{\eta_2 r_1(0) + \eta_1 r_2(0)}{(\eta_1+\eta_2)} \\ & + \frac{\eta_k^2(l_1^2(0) + l_2^2(0) - 2Q(0))}{(\eta_1+\eta_2)^2}\exp(-(\eta_1+\eta_2)(2-(\eta_1+\eta_2))t) \\ & + (-1)^k \frac{2\eta_k(\eta_2 l_1(0) - \eta_1 l_2(0) - (\eta_2-\eta_1)Q(0))}{(\eta_1+\eta_2)^2}\exp(-(\eta_1+\eta_2)t) + \frac{\eta_2^2 l_1^2(0) + \eta_1^2 l_2^2(0) + 2\eta_1\eta_2 Q(0)}{(\eta_1+\eta_2)^2} \Bigg\} \end{aligned} \qquad (29)$$

4 Results

When the step sizes of two students are the same, the mutual learning asymptotically converges to the average weight vector of two students [12]. In this section, we analyze the asymptotic property of mutual learning in the case of different step sizes, and then discuss the relationship between mutual learning and ensemble learning.

4.1 Effect of step size on the asymptotic property of mutual learning

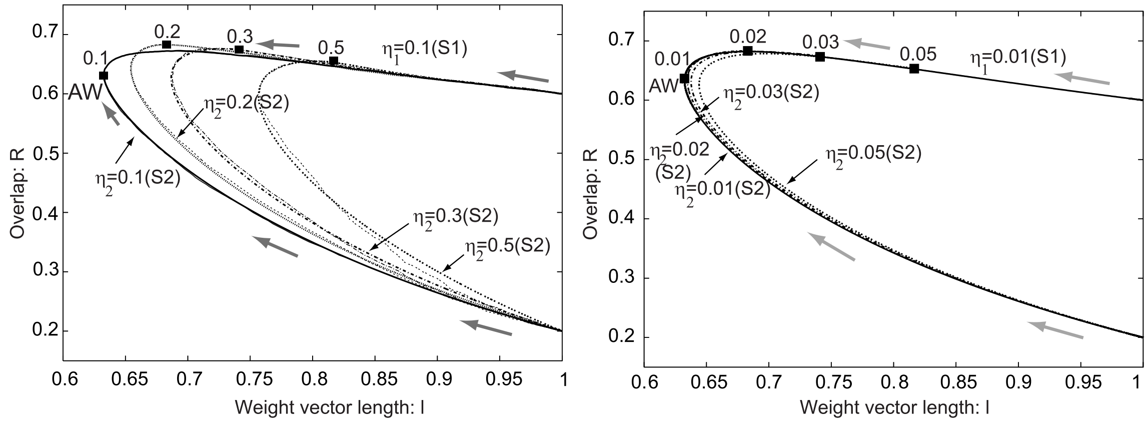
We analyze the effect of the learning step size on the asymptotic property of mutual learning. Two students use different learning step sizes. For this purpose, we use computer simulations.

Figure 2 shows trajectories of the student weight vectors when the initial overlaps between the latent teacher and the students were inhomogeneous: (a) shows the results obtained through setting the learning step size of student 1 (η_1) to 0.1(fixed), and setting the learning step size of student 2 (η_2) to 0.1, 0.2, 0.3, or 0.5; (b) shows the results obtained through setting the learning step size η_1 to 0.01(fixed), and setting η_2 to 0.01, 0.02, 0.03, or 0.05. In these figures, the horizontal axis shows the length of the student weight vector l_k, and the vertical axis shows the overlap R_k. The initial conditions were $l_1(0) = l_2(0) = 1$, $R_1(0) = 0.6$, $R_2(0) = 0.2$, and $q(0) = -0.2$. The theoretical results obtained using Eqs. (13), (14), and (21) are shown as thick lines, and the results obtained through computer simulations for $N = 10000$ are shown as thin lines. The upper lines show trajectories of the weight vector of student 1, and the lower lines show trajectories of the weight vector of student 2. The symbols of black rectangles show convergence points of trajectories of the student weight vectors. The numbers above the symbols show the learning step sizes of student 2.

When the learning step sizes η_1 and η_2 were the same, student 1 started at $l_1(0) = 1$ and $R_1(0) = 0.6$, and converged to the average weight vector of the initial student vectors denoted by $\boldsymbol{AW}$. Student 2 started at $l_2(0) = 1$ and $R_2(0) = 0.2$, and also converged to the average weight vector denoted by $\boldsymbol{AW}$ when using the same learning step sizes.

When the learning step sizes η_1 and η_2 were not the same, the convergence points were changed by using a different step size η_2 of $0.2, 0.3$, or 0.5 as shown in Fig. 2(a). As in Fig. 2(a), Fig. 2(b) shows that the convergence points were changed by using a different step size η_2 of $0.02, 0.03$, or 0.05. Note that the convergence points for the same ratio of the learning step size tend to be the same. Thus, we pay attention to the effect of the ratio of learning step sizes η_2/η_1 in the mutual learning.

Figure 3 shows the learning step size dependence of the generalization error. The learning step size of student 1 was 0.1 or 0.01(fixed), and that of student 2 was changed as shown in the figure. The horizontal axis shows the ratio of learning step sizes η_2/η_1, and the vertical axis shows the asymptotic property of the generalization error ϵ_g. The asymptotic property of the generalization error



(a) $\eta_1 = 0.1$

(b) $\eta_1 = 0.01$

Figure 2: Trajectories of student weight vector for the inhomogeneous case. The initial conditions were $l(0) = 1$, $R_1(0) = 0.6$, $R_2(0) = 0.2$, and $q(0) = -0.2$. (a) Results of setting the learning step size to $\eta_1 = 0.1$(fixed) and $\eta_2 = 0.1, 0.2, 0.3$, or 0.5. (b) Results of setting the learning step size to $\eta_1 = 0.01$(fixed) and $\eta_2 = 0.01, 0.02, 0.03$, or 0.05.

is obtained using Eq. (29) for the case of $t \to \infty$. The results show that the asymptotic property of the generalization error was minimized when the ratio η_2/η_1 was 2. Consequently, the asymptotic property of the generalization error can be minimized by using the optimal ratio of learning step sizes. Next, we will obtain this optimal ratio of learning step sizes that minimizes the asymptotic property of the generalization error.

4.2 Optimization of the asymptotic property of the generalization error

We now analyze the asymptotic property of the generalization error based on the ratio of learning step sizes, and then we obtain the optimum ratio of learning step sizes η_2/η_1 that minimizes the asymptotic property of the generalization error.

The asymptotic property of the order parameters is obtained by substituting $t \to \infty$ into Eqs. (13), (14), and (21):

$$l_1^2(\infty) = l_2^2(\infty) = Q(\infty) = \frac{\eta_2^2 l_1^2(0) + \eta_1^2 l_2(0) + 2\eta_1\eta_2 Q(0)}{(\eta_1 + \eta_2)^2}, \tag{30}$$

$$r_1(\infty) = r_2(\infty) = \frac{\eta_2}{\eta_1 + \eta_2} r_1(0) + \frac{\eta_1}{\eta_1 + \eta_2} r_2(0). \tag{31}$$

The above equations show that the mutual learning converges to the internal dividing point of the initial student weight vectors. Using Eqs. (30) and (31),

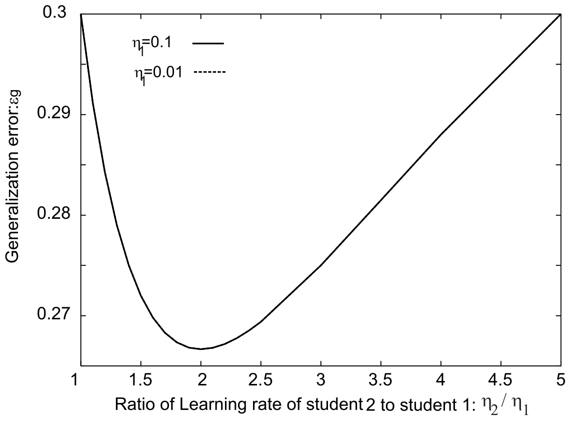


Figure 3: Relation between learning step size and generalization error. The learning step size of student 1 was 0.1 or 0.01(fixed), and that of student 2 was changed. The generalization error is minimized when the ratio of the learning step size is two for both cases. The optimum ratio is independent of the size of the learning step size.

we can obtain the asymptotic property of the generalization error:

$$\epsilon_g(\infty) = \frac{1}{2}\left\{1 - 2\frac{\eta_2 r_1(0) + \eta_1 r_2(0)}{\eta_1 + \eta_2} + \frac{\eta_2^2 l_1^2(0) + \eta_1^2 l_2(0) + 2\eta_1\eta_2 Q(0)}{(\eta_1 + \eta_2)^2}\right\} \quad (32)$$

We rewrite the generalization error by replacing the ratio η_2/η_1 with α:

$$\epsilon_g(\infty) = \frac{1}{2}\left\{1 - 2\frac{\alpha r_1(0) + r_2(0)}{\alpha + 1} + \frac{\alpha^2 l_1^2(0) + l_2^2(0) + 2\alpha Q(0)}{(\alpha + 1)^2}\right\}. \quad (33)$$

When the generalization error is minimized, $\partial\epsilon_g(\infty)/\partial\alpha = 0$ is satisfied, so

$$\frac{\partial\epsilon_g}{\partial\alpha} = \frac{2\alpha l_1^2(0) + 2Q(0)}{(\alpha+1)^2} - \frac{2(\alpha^2 l_1^2(0) + l_2^2(0) + 2\alpha Q(0))}{(\alpha+1)^2} + \frac{2(\alpha r_1(0) + r_2(0))}{(\alpha+1)^2} - \frac{2r_1(0)}{\alpha+1} = 0 \quad (34)$$

Solving Eq. (34), we obtain α^{opt} as

$$\alpha^{opt} = \frac{l_2^2(0) - Q(0) + r_1(0) - r_2(0)}{l_1^2(0) - Q(0) - r_1(0) + r_2(0)}. \quad (35)$$

Therefore, the optimum ratio of the learning step size is obtained through Eq. (35). The optimum asymptotic property of the generalization error is obtained by substituting Eq. (35) into Eq. (33):

$$\epsilon_g^{opt}(\infty) = \frac{1}{2}\left\{1 - 2(\kappa r_1(0) + (1-\kappa)r_2(0)) + \kappa^2 l_1^2(0) + (1-\kappa)^2 l_2^2(0) + 2\kappa(1-\kappa)Q(0)\right\}. \quad (36)$$

Here, κ is defined as $\kappa = \alpha^{opt}/(1 + \alpha^{opt})$.

On the other hand, we can consider the linear combination of the initial weight vectors of the students — that is, $\boldsymbol{J} = C\boldsymbol{J}^1(0) + (1 - C)\boldsymbol{J}^2(0)$ — and minimize the generalization error by C. This is an ensemble learning with two students, so from the appendix, the optimum C^* that minimizes the generalization error is

$$C^* = \frac{l_2^2(0) - Q(0) + r_1(0) - r_2(0)}{l_1^2(0) + l_2^2(0) - 2Q(0)}. \tag{37}$$

Therefore, the optimum ratio $C^*/(1 - C^*)$ is obtained as

$$\frac{C^*}{1 - C^*} = \frac{l_2^2(0) - Q(0) + r_1(0) - r_2(0)}{l_1^2(0) - Q(0) - r_1(0) + r_2(0)} = \frac{\eta_2^{opt}}{\eta_1^{opt}}, \tag{38}$$

and it is shown that the optimum ratio of the learning step size of mutual learning $\alpha^{opt} = \eta_2^{opt}/\eta_1^{opt}$ is equal to that of the optimum linear combination of the initial weight vectors $C^*/(1 - C^*)$. Consequently, mutual learning using an optimum ratio of learning step sizes converges to the optimum ensemble learning that is the linear combination of the initial student vectors.

5 Conclusion

We have proposed an optimization method for mutual learning by means of minimizing the asymptotic property of the generalization error within the framework of on-line learning. We first formulated mutual learning with a latent teacher, and then derived the differential equations of order parameters that depict the learning process. The order parameters of mutual learning are the length of the student weight vector l_k and the overlap between students q. To depict the relationship between the latent teacher and the students, we introduced the order parameter R_k. We derived these differential equations using statistical mechanics methods and solved them analytically. After that, we obtained the dynamics of the generalization error using these order parameters.

Next, we used the theoretical results to analyze the relationship between the asymptotic property of the mutual learning and the learning step size of the students. From the results, we found that the asymptotic property of the mutual learning related to the ratio of the learning step sizes of two students, and was not related to the learning step size itself. We obtained the optimum ratio of the learning step size which minimizes the generalization error analytically. We also showed that the optimum ratio of the learning step sizes of the mutual learning is equal to the inverse of the ratio of optimum weights for an average of the linear combination of initial student weight vectors. We conclude that the integration mechanism of ensemble learning can be mimicked through mutual learning by introducing the interaction between students. Our future work will include analysis of the mutual learning with non-linear perceptrons.

Acknowledgment

We would like to thank Masato Okada (The University of Tokyo) and Seiji Miyoshi (Kobe City College of Technology) for their useful discussions. Part of this study has been supported by a Grant-in-Aid for Scientific Research (C) No. 16500146.

References

[1] L. Breiman, Bagging predictors, *Machine Learning*, vol. 24, pp. 123-140 (1996).

[2] Y. Freund and R. E. Shapire, J. Comput. Syst. Sci. **55** (1997) 119.

[3] On-line Learning in Neural Networks, ed. D. Saad (Cambridge University Press, Oxford, 1998).

[4] A. Krogh and P. Sollich, Phys. Rev. E, **55** (1997) 811.

[5] K. Hara and M. Okada, Neural Networks, **17** (2004) 215.

[6] K. Hara and M. Okada, J. Phys. Soc. Jpn. **74** (2005) 2966.

[7] S. Miyoshi, K. Hara, and M. Okada, Phys. Rev. E, **71** (2005) 036116.

[8] A. Lazarevic and Z. Obradivic, Distributed and parallel databases, vol.11, pp. 203 (2002).

[9] Klein, E., et. al., Proc. Neural Inf. Pro. Sys. (2004).

[10] R. Metzler, W. Kinzel, and I. Kanter: Phys. Rev. E 62 (2000) 2555.

[11] R. Mislovaty, E. Klein, I. Kanter, and W. Kinzel: Phys. Rev. Lett. 91 (2003) 118701.

[12] Hara K. and M. Okada, J. Phys. Soc. Jpn. **76** (2007) 014001.

A Ensemble learning

Ensemble learning is a learning method using many weak learning machines to improve upon the performance of a single weak learning machine[1, 2, 8]. Students learn from the teacher individually, and then an ensemble output is calculated by integrating the students' outputs. Because many students are used, ensemble learning is effective when the students differ from each other. Therefore, we assume that the overlap(direction cosine) between the kth student and the k'th student $q_{kk'}$ is not one. The ensemble output of the student networks $\overline{u}$ is given by the weighted average of each student output using the weights for averaging C_k:

$$\overline{u} = \sum_{k=1}^{K} C_k u_k = \sum_{k=1}^{K} C_k \Big(\boldsymbol{J}^k \cdot \boldsymbol{x} \Big) \tag{39}$$

Here, the number of students is K and we assume $\sum_{k=1}^{K} C_k = 1$. In the following, we assume that the number of students is two. We use linear perceptrons as the students, so the average output of the two students is equal to the output of a perceptron having the average of the two student weight vectors. The weighted average of the two student weight vectors $\boldsymbol{J}^E$ is defined as follows[12].

$$\boldsymbol{J}^E = C_k \boldsymbol{J}^k + C_{k'} \boldsymbol{J}^{k'} = C \boldsymbol{J}^k + (1-C) \boldsymbol{J}^{k'} \tag{40}$$

Here, we rewrite C_k as C and $C_{k'}$ as $1-C$ from $C_k + C_{k'} = 1$. From this equation, ensemble learning can be viewed as the linear combination of the two student weight vectors. Note that ensemble learning is a static process, so there is no dynamical property. The length of the weight vector l^E and the overlap r^E are given by

$$(l^E)^2 = C^2 l_k^2 + (1-C)^2 l_{k'}^2 + 2C(1-C)Q \tag{41}$$

$$r^E = C r_k + (1-C) r_{k'} \tag{42}$$

The generalization error of ensemble output ϵ_g^E is given by substituting Eqs. (41) and (42) into Eq. (28):

$$\begin{aligned} \epsilon_g^E &= \frac{1}{2}\left(1 - 2r^E + (l^E)^2\right) \\ &= \frac{1}{2}\Big\{ 1 - 2(C r_k + (1-C) r_{k'}) + C^2 l_k^2 + (1-C)^2 l_{k'}^2 + 2C(1-C)Q \Big\}. \end{aligned} \tag{43}$$

If the optimum weight for average C^* satisfies the condition of $\partial \epsilon_g^* / \partial C^* = 0$, we obtain

$$C^* = \frac{l_{k'}^2 - Q + r_k - r_{k'}}{l_k^2 + l_{k'}^2 - 2Q} \tag{44}$$

When the student weight vector length $l_k = l_{k'} = l$ and the overlap between the students $r_k = r_{k'} = r$, from Eq. (44) we obtain $C^* = (1 - C^*) = 1/2$ and the simple average of the two students is the optimum ensemble output.